It's the genes! EST access to human genome content

David Gerhold and C. Thomas Caskey

Summary

ESTs or 'expressed sequence tags' are DNA sequences read from both ends of expressed gene fragments. The Merck-WashU EST Project and several other public EST projects are being performed to rapidly discover the complement of human genes, and make them easily accessible. These ESTs are widely used to discover novel members of gene families, to map genes to chromosomes as 'sequence-tagged sites' (STSs), and to identify mutations leading to heritable diseases. Informatic strategies for querying the EST databases are discussed, as well as the strengths and weaknesses of the EST data. There is a compelling need to build on the informatic synthesis of human gene data, and to devise facile methods for determining gene functions.

Accepted
18 October 1996

Introduction

Our understanding of molecular biology is built upon a traditional foundation. Start with a defined biological problem, find the genes that determine the phenotype of interest, and incorporate techniques from cell biology, biochemistry, genetics, etc. to elucidate the biochemical pathway. This paradigm has allowed us to answer simple questions in man, and complex questions in model eukaryotes and bacteria. The genome initiative, however, has taught us that we can now address comprehensive questions about the human genome. Which genes are expressed in each cell type? Are genes only expressed where they are needed? What is the importance of a genes' location and context within the genome? What are the transcriptional and translational elements of each gene? How and why are genes alternatively spliced? What is the extent of genetic variation between individuals and how does it influence human health, disease and behavior? While some of these questions await a complete sequence and analysis of the human genome, many can be addressed by collectively analyzing the <5% of the genome that is transcribed and translated into protein.

Development of ESTs/STSs

The central dogma of molecular biology holds that genes encoded in DNA are copied into messenger RNA (mRNA), which is then translated into functional proteins. Molecular biologists typically study expressed genes by isolating mRNAs by means of their 3'poly(A) tails, and copying them into complementary DNA, or cDNA. These cDNAs represent fragments of individual genes, which can be 'cloned' into DNA circles called plasmids, and replicated many times in *E. coli*.

In the 1980s the advent of high-throughput automated sequencing made it possible randomly to select many cDNA clones from plasmid cDNA libraries and to determine the DNA sequence of several hundred bases from both ends. These short DNA sequences are called 'Expressed Sequence Tags', or ESTs, and the position of each gene or other DNA marker on a physical chromosome map is called a Sequence Tagged Site, or STS. Since ESTs and STSs are sequence-based, each is amenable to PCR amplification, a powerful tool for searching and characterizing genes. The EST sequence is sufficient to identify known genes, and to glimpse the biochemical functions of many novel genes.

As an example, Merck-WashU ESTs representing tumor suppressor gene p53 are depicted in Fig. 1. In the case of p53, the most common mutant gene in neoplasias, 5'EST:H61357 represents a partial coding region and 3'EST:H62385 represents a 3' noncoding sequence from the same clone. 5'EST:T80132 does not overlap 5'EST:H61357, but is readily assigned to the same gene by identity in the 3'ESTs from the two clones

History of the EST approach

Sequencing of randomly selected hepatic cDNAs demonstrated the utility of DNA sequence-to-gene function relationships as early as 1983[1].

The EST approach was described in 1992 almost simultaneously by Sikela[2] and Matsubara[3], and was pursued on a

©ICSU Press 1996 pp 973-981 DIO1048